AI & ML

SESSION 20

MODULE 5

Inclusys Org Foundation
Empowering the Neurodivergent

Session Outline

- Curse of Dimensionality
- Dimensionality Reduction
- Feature Selection
- Matrix Factorization based DR
- Preliminaries
- PCA

Inclusys Org Foundation
Empowering the Neurodivergent

Can you rank the models ?

Model 1
D = 10
Accuracy a1

Model 2
D = 50
Accuracy a2

Model 3
D = 100
Accuracy a3

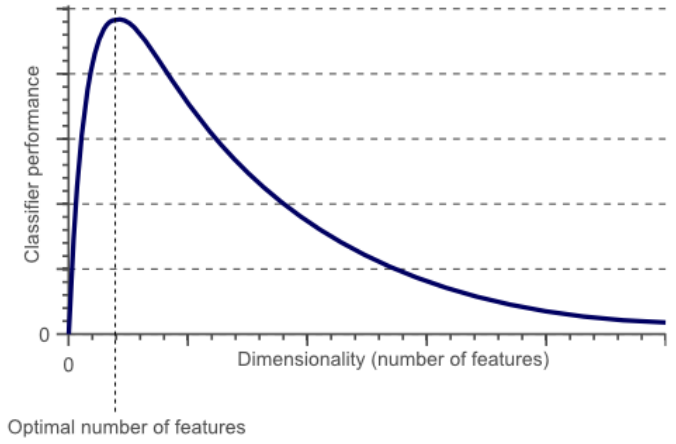
Model 5
D = 200
Accuracy a4

Model 5
D = 500
Accuracy a5

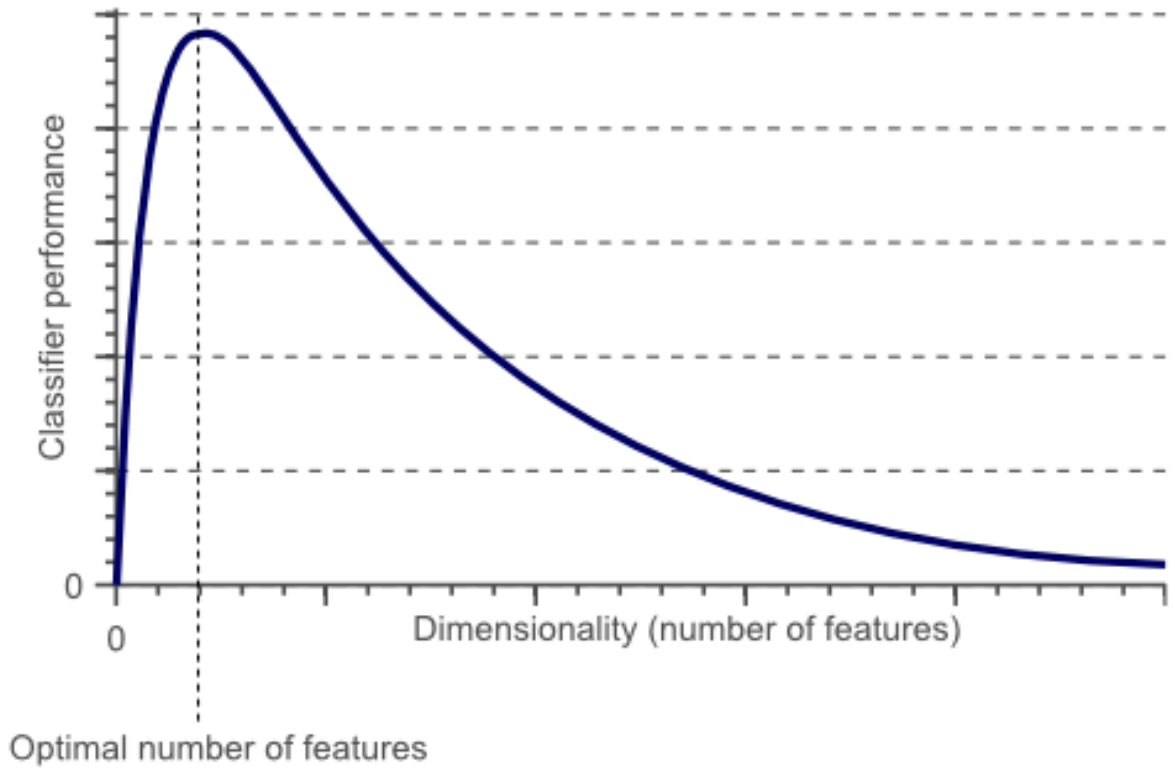
Inclusys Org Foundation
Empowering the Neurodivergent

Can you rank the models ?

Model 1 D = 10 Accuracy a1	Model 2 D = 50 Accuracy a2	Model 3 D = 100 Accuracy a3	Model 5 D = 200 Accuracy a4	Model 5 D = 500 Accuracy a5



Hughes Phenomenon



Curse of Dimensionality

- Term coined by Bellman in 1961.
- The sample size needed to estimate a function of several variables to a given degree of accuracy (i.e. to get a reasonably low-variance estimate) grows exponentially with the number of variables.
- as the number of attributes or the dimensions increases, the number of training samples required to generalize a model also increases phenomenally.
- high-dimensional spaces are inherently sparse.
- Overfitting models.

Combating CoD

A way to avoid the curse of the dimensionality is to reduce the input dimension of the function to be estimated.

Dimensionality Reduction

- Dimensionality – no of input variables.
- Large numbers of input features can cause poor performance for machine learning algorithms.
- Dimensionality reduction is a general field of study concerned with reducing the number of input features.
- Dimensionality reduction methods include feature selection, linear algebra methods, projection methods, and autoencoders.

When to do

- Performed after data cleaning and data scaling
- Before training a predictive model.

Any dimensionality reduction performed on training data must also be performed on new data, such as a test dataset, validation dataset, and data when making a prediction with the final model.

Common Data Preparation Tasks

- Data Cleaning
- Feature Selection
- Data Transforms
- Feature Engineering
- Dimensionality Reduction

Inclusys Org Foundation
Empowering the Neurodivergent

Common Data Preparation Tasks

- Data Cleaning
- **Feature Selection**
- Data Transforms
- Feature Engineering
- **Dimensionality Reduction**

- **Reduce computational cost of modeling**
- **Improve model performance**

Techniques for Dimensionality Reduction

Feature Selection Methods

Inclusys Org Foundation
Empowering the Neurodivergent

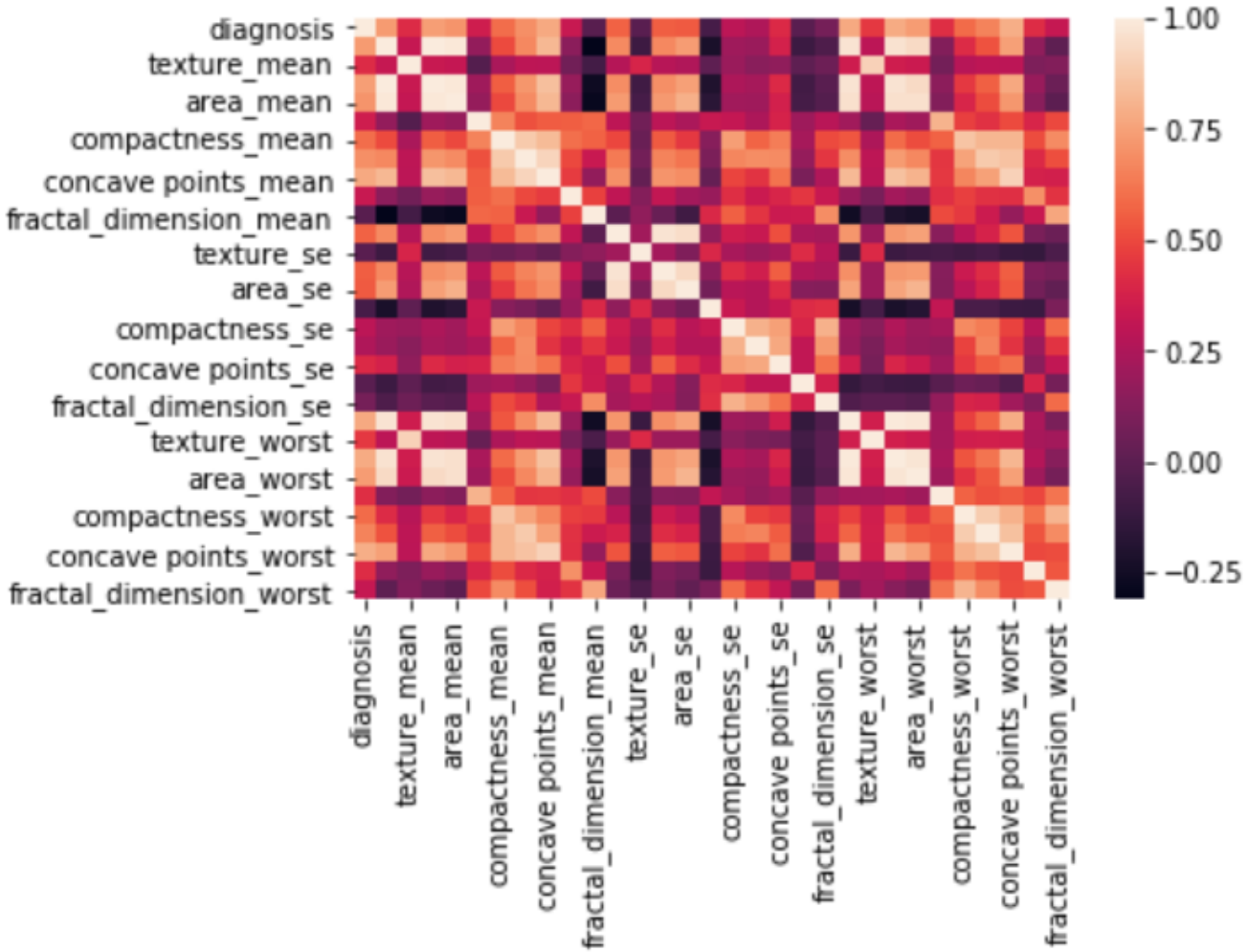
Feature Selection

Process of reducing the number of input variables when developing a predictive model.

Already seen

- Delete columns with id, names.
- Delete columns with constant values.
- Delete columns that exhibit low variance.

Correlation



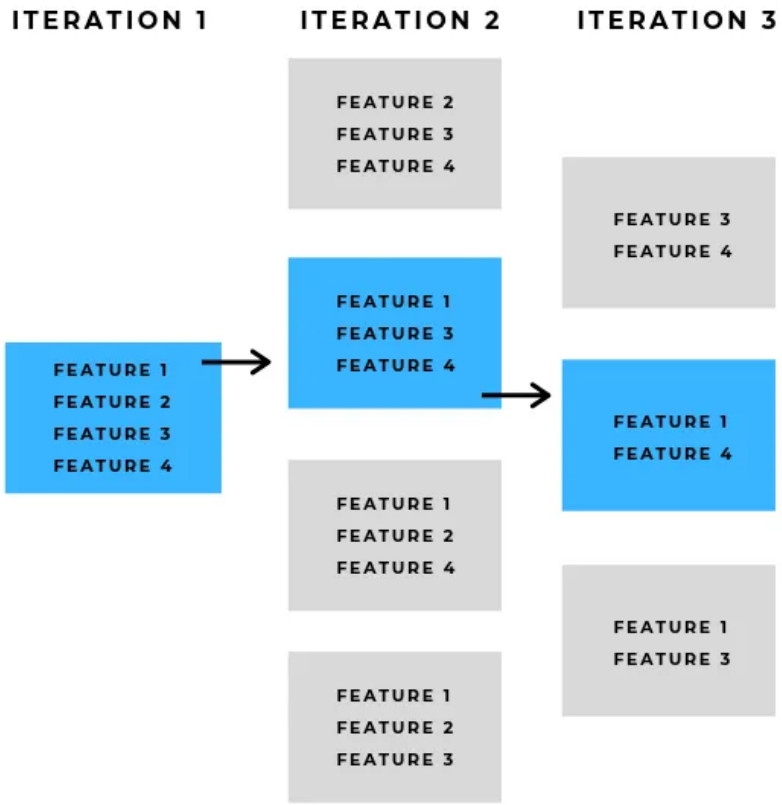
Inclusys Org Foundation
Empowering the Neurodivergent

Score based Selection

Selecting K best features

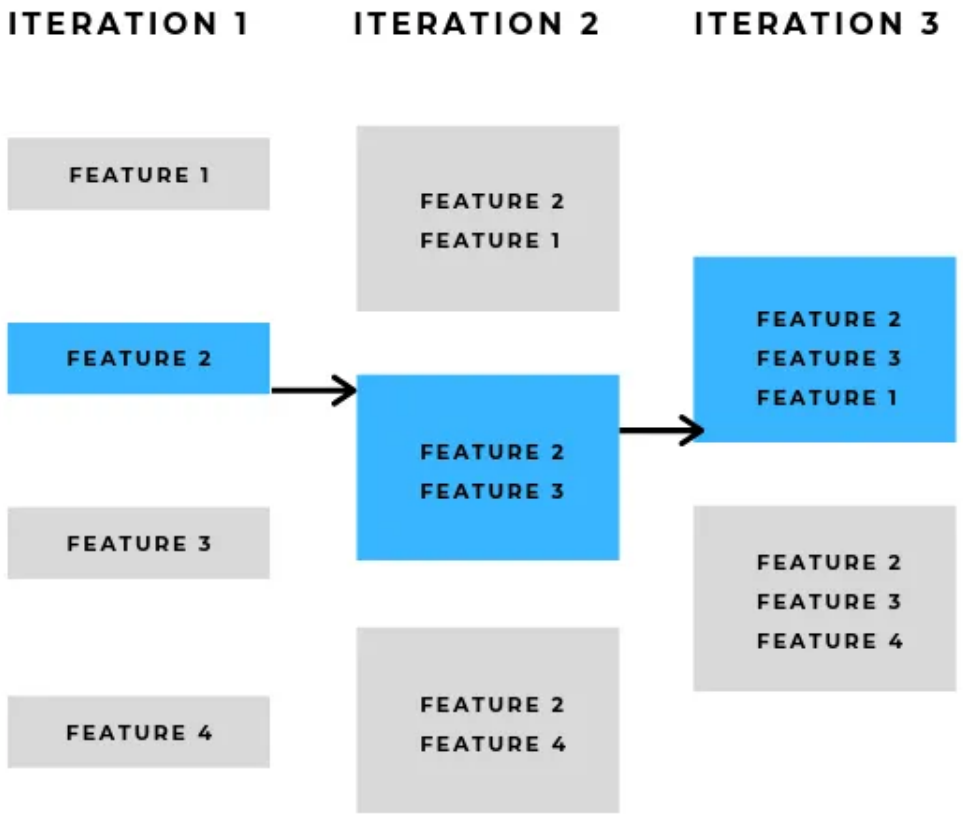
sklearn.feature_selection.SelectKBest — scikit-learn 1.2.1 documentation

Recursive Feature Elimination



https://medium.com/analytics-vidhya/feature-selection-methods-for-data-science-just-a-few-fca3086eb445

Sequential Feature Selection

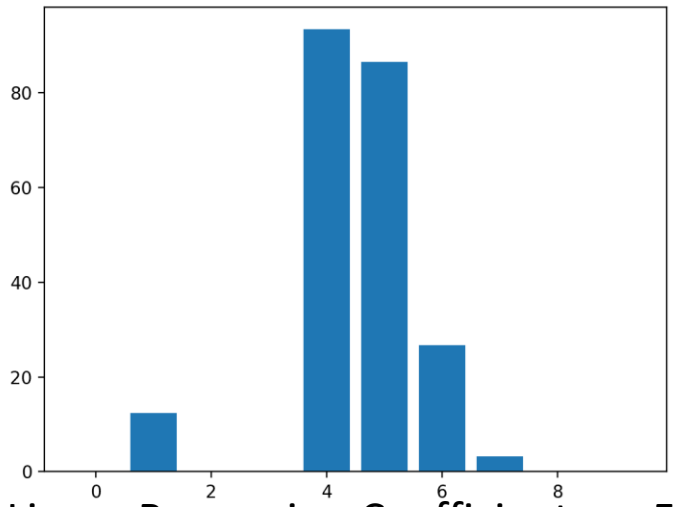


https://medium.com/analytics-vidhya/feature-selection-methods-for-data-science-just-a-few-fca3086eb445

Feature Importance based selection

Linear machine learning algorithms fit a model where the prediction is the weighted sum of the input values.

All of these algorithms find a set of coefficients to use in the weighted sum in order to make a prediction. These coefficients can be used directly as a crude type of feature importance score.



Linear Regression Coefficients as Feature Importance Scores

Feature: 0, Score: 0.00000
Feature: 1, Score: 12.44483
Feature: 2, Score: -0.00000
Feature: 3, Score: -0.00000
Feature: 4, Score: 93.32225
Feature: 5, Score: 86.50811
Feature: 6, Score: 26.74607
Feature: 7, Score: 3.28535
Feature: 8, Score: -0.00000
Feature: 9, Score: 0.00000

Matrix Factorization

Matrix Factorization

Matrix factorization methods can be used to reduce a dataset matrix into its constituent parts.

Examples include the eigen decomposition and singular value decomposition.

The parts can then be ranked and a subset of those parts can be selected that best captures the salient structure of the matrix that can be used to represent the dataset.

The most common method for ranking the components is principal components analysis, or PCA for short.

Preliminaries

Variance

A measure of the spread of the data in a data set with mean, $\bar{x}$

$$\sigma^2 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{(n-1)}$$

Variance is claimed to be the original statistical measure of spread of data.

Covariance

Variance – measure of the deviation from the mean for points in one dimension, e.g., heights

Covariance – a measure of how much each of the dimensions varies from the mean with respect to each other.

Covariance is measured between 2 dimensions to see if there is a relationship between the 2 dimensions, e.g., number of hours studied and grade obtained.

The covariance between one dimension and itself is the variance

$$\text{var}(X) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})}{(n-1)}$$

$$\text{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{(n-1)}$$

Covariance Matrix

Representing covariance among dimensions as a matrix, e.g., for 3 dimensions:

$$C = \begin{bmatrix} \mathrm{cov}(X,X) & \mathrm{cov}(X,Y) & \mathrm{cov}(X,Z) \\ \mathrm{cov}(Y,X) & \mathrm{cov}(Y,Y) & \mathrm{cov}(Y,Z) \\ \mathrm{cov}(Z,X) & \mathrm{cov}(Z,Y) & \mathrm{cov}(Z,Z) \end{bmatrix}$$

Properties:

- Diagonal: variances of the variables
- cov(X,Y)=cov(Y,X), hence matrix is symmetrical about the diagonal (upper triangular)
- m-dimensional data will result in mxm covariance matrix

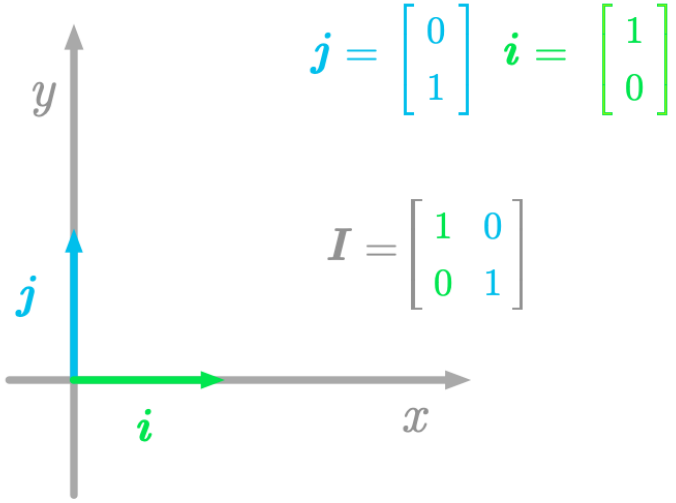
Linear Independence

A set of n-dimensional vectors xi Є Rn, are said to be linearly independent if none of them can be written as a linear combination of the others.

$$c_1x_1 + c_2x_2 + \ldots + c_kx_k = 0$$

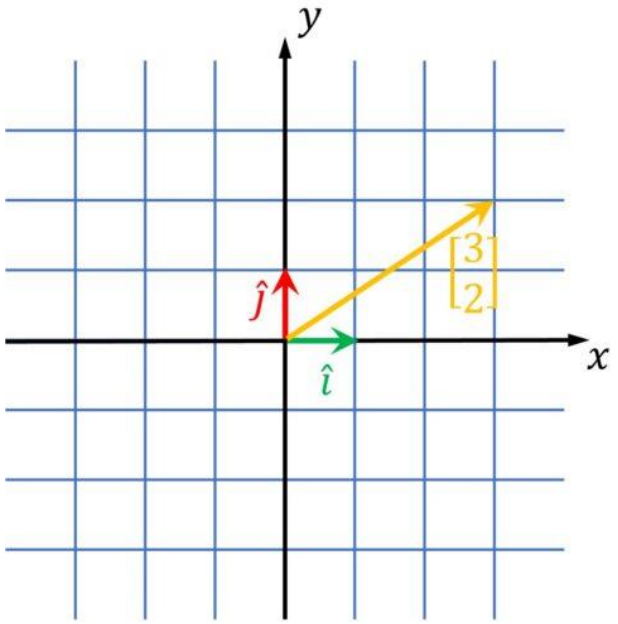
$$iff \qquad c_1 = c_2 = \ldots = c_k = 0$$

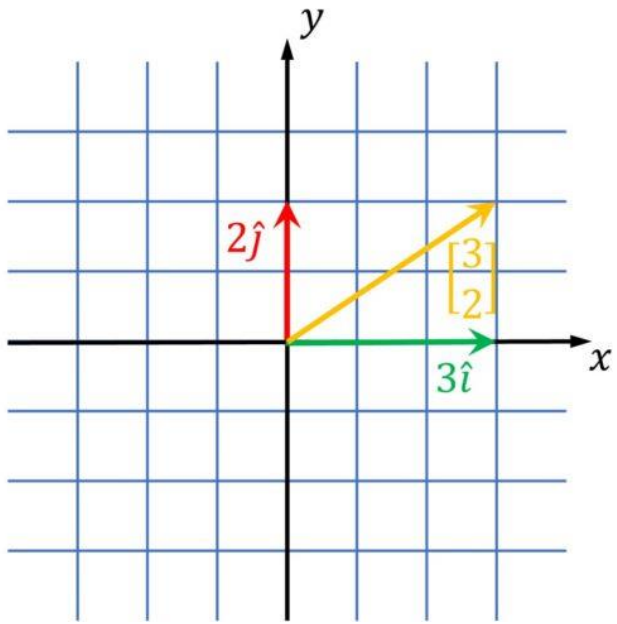
Basis



$$\boldsymbol{j} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad \boldsymbol{i} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$\boldsymbol{I} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$





?

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \times \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

Example

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \times \begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 12 \\ 8 \end{bmatrix} = 4 \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

$$A \times v = \lambda v$$

Eigen values and Eigen vectors

$$A \times v = \lambda v$$

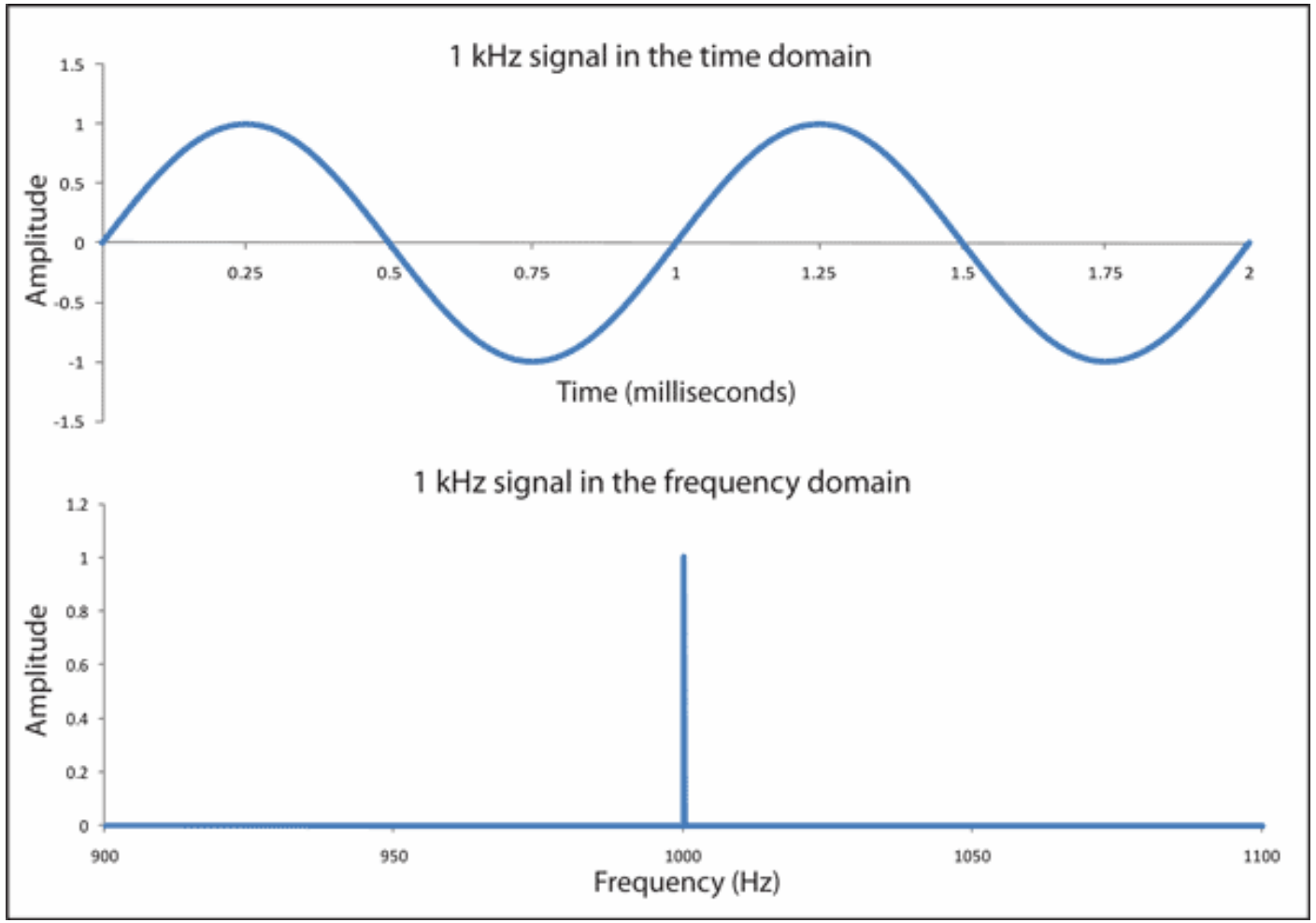
A: m x m matrix

v: m x 1 non-zero vector

λ: scalar

Any value of λ for which this equation has a solution is called the eigenvalue of A and the vector v which corresponds to this value is called the eigenvector of A.

Principal Component Analysis


1 kHz signal in the time domain
Amplitude
1.5
1
0.5
0
-0.5
-1
-1.5
0.25
0.5
0.75
1
1.25
1.5
1.75
2
Time (milliseconds)
1 kHz signal in the frequency domain
Amplitude
1.2
1
0.8
0.6
0.4
0.2
0
900
950
1000
1050
1100
Frequency (Hz)

Change of Basis

Let X and Y be m x n matrices related by a linear transformation P.

X is the original recorded data set and Y is a re-representation of that data set.

$$PX = Y$$

Let's define;

- pi are the rows of P.
- xi are the columns of X.
- yi are the columns of Y.

PX =Y

- P is a matrix that transforms X into Y.
- Geometrically, P is a rotation and a stretch (scaling) which again transforms X into Y.
- The rows of P, {p1, p2, …,pm} are a set of new basis vectors for expressing the columns of X.

➢ *Changing the basis doesn't change the data – only its representation.*

➢ *Changing the basis is actually projecting the data vectors on the basis vectors.*

Change of basis problem

- Assuming linearity, the problem now is to find the appropriate change of basis.
- Now,
  - What is the best way to re-express X?
  - What is the good choice of basis P?

Principal Component Analysis

a statistical procedure to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables.

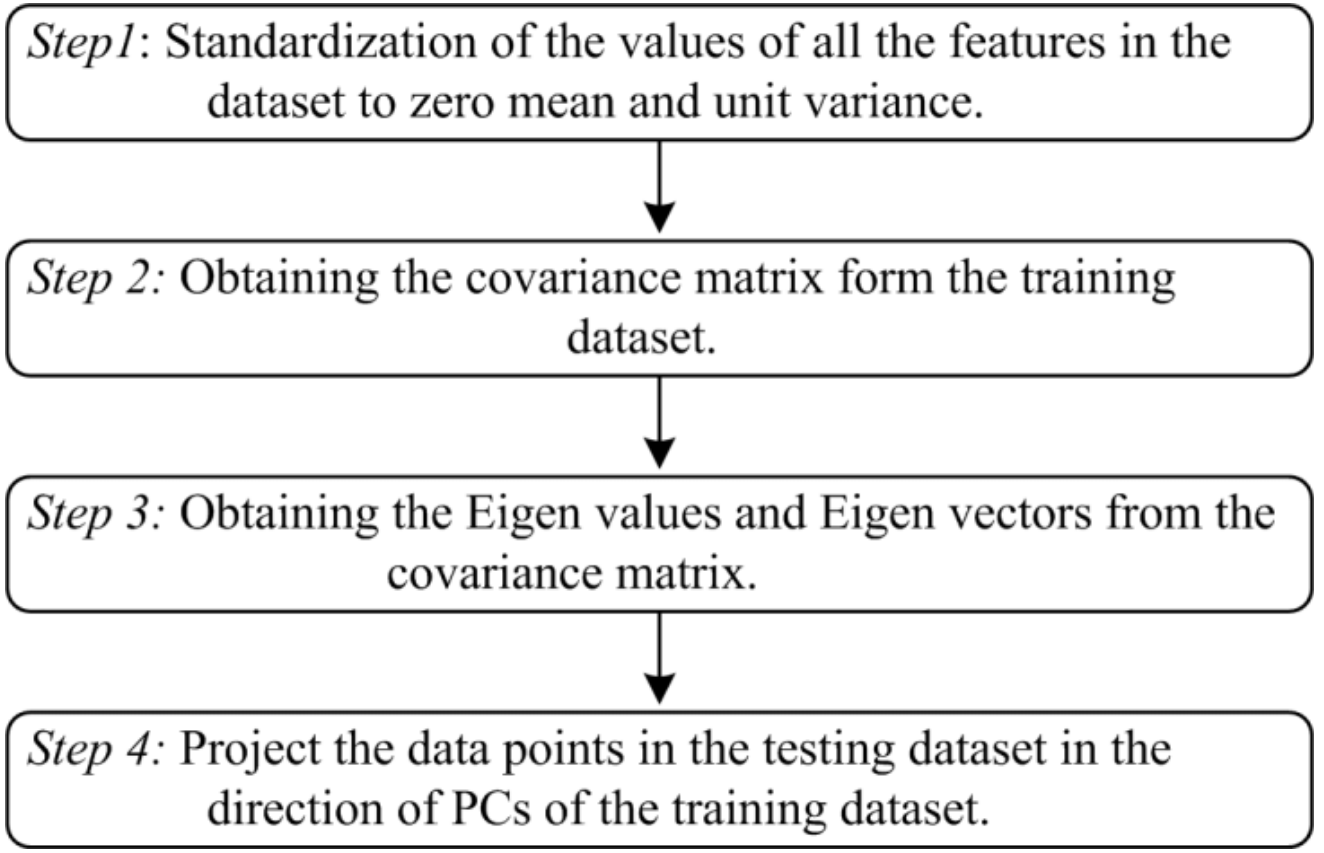
Each of the principal components is chosen in such a way so that it would describe most of them still available variance .

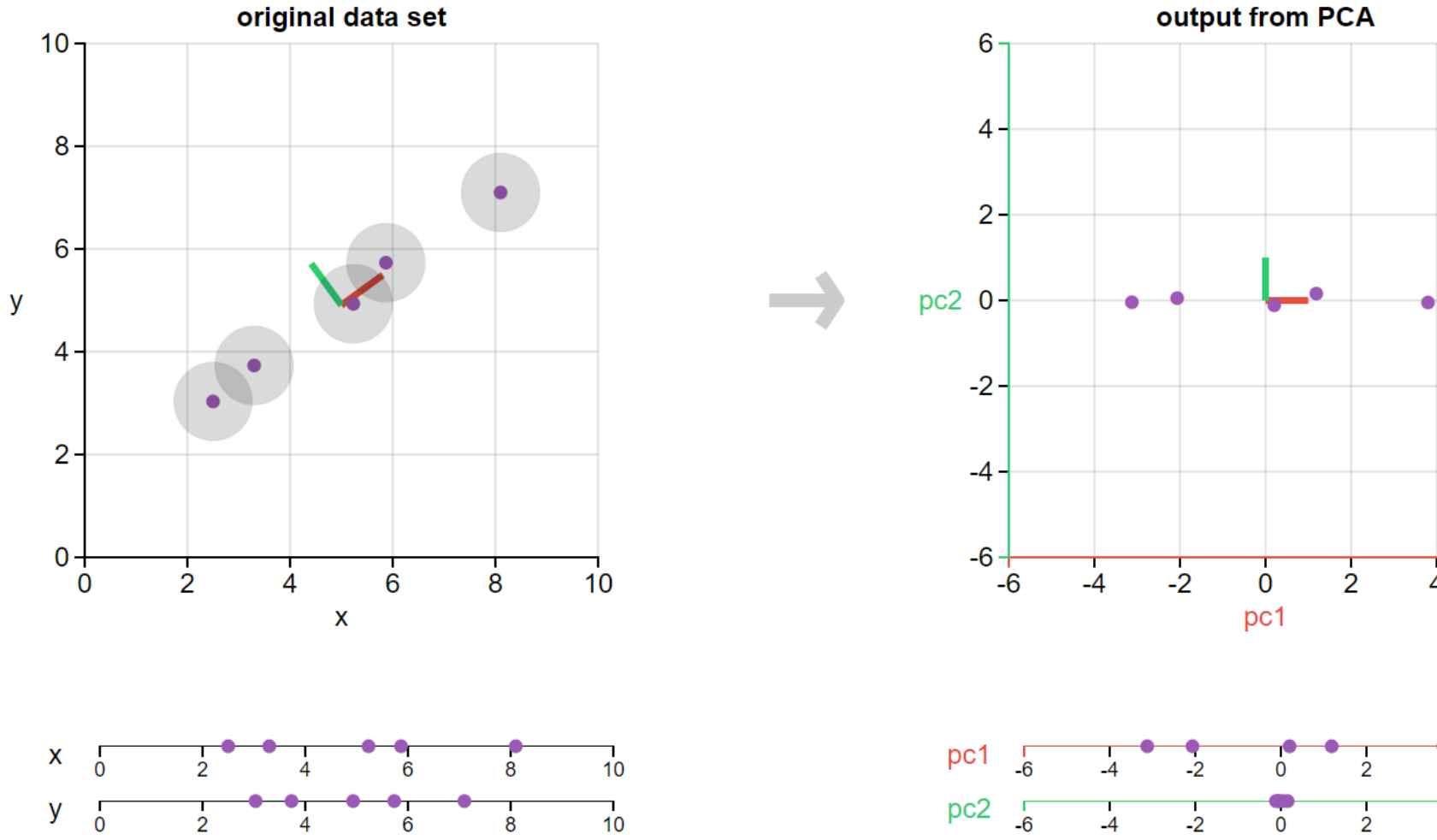
All these principal components are orthogonal to each other.

In all principal components first principal component has a maximum variance.

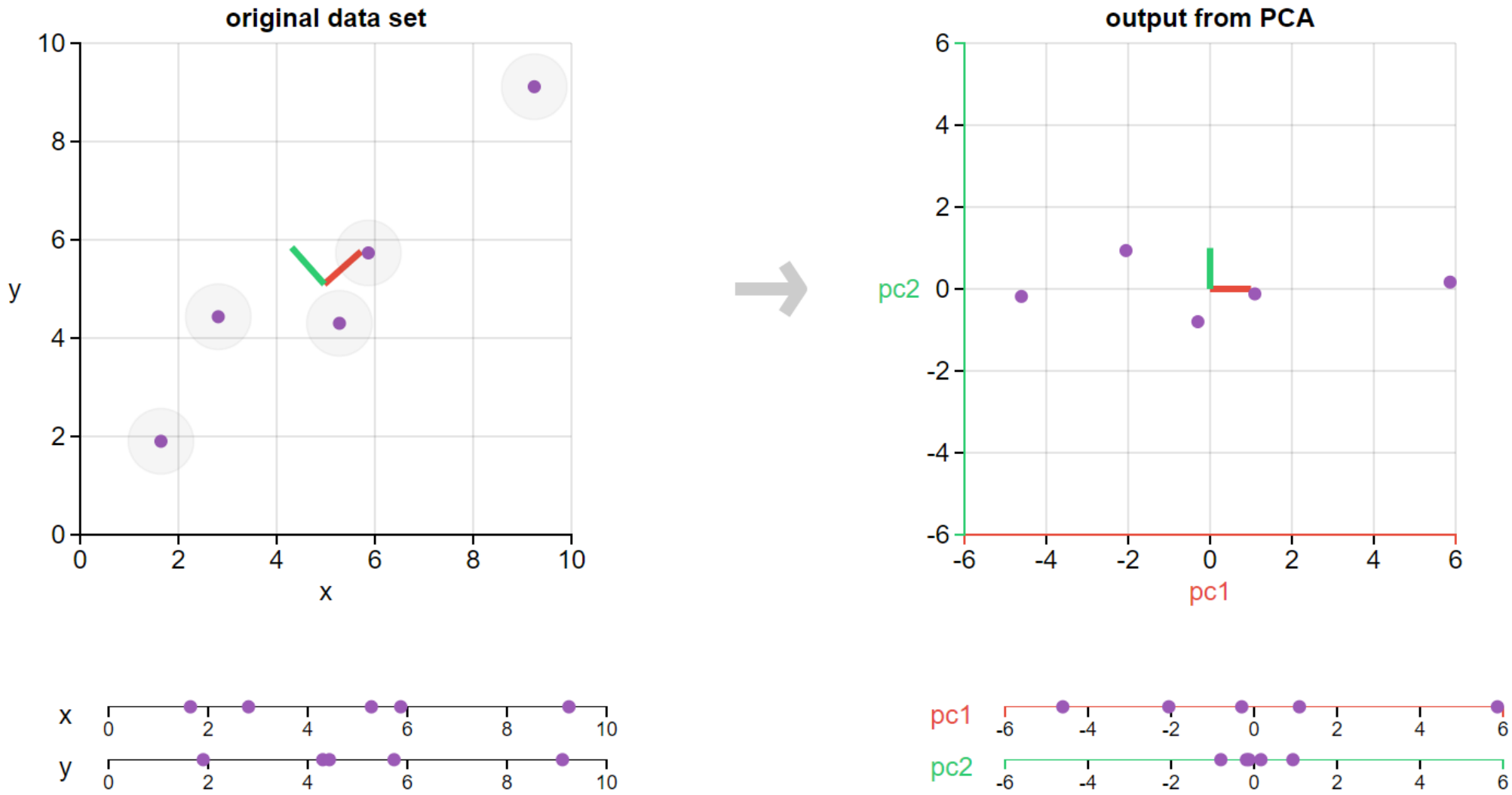
These are basically performed on a square symmetric matrix.

Steps in PCA





[Principal Component Analysis explained visually (setosa.io)]


original data set
y
x
0
2
4
6
8
10
output from PCA
pc2
pc1
-6
-4
-2
0
2
4
6
x
y
pc1
pc2

PCA on a dataset

Thank You!